Your Second RecSys

Елисова Ирина
ML Teamlead
MTC BigData

Двухэтапная модель

План

Мотивация

Схема двухэтапной модели

Модели 1 и 2 этапа

Схема валидации

Сбор обучающей выборки

Метрики качества

МТС ods.ai

Простые модели. Что дальше?

implicit

SVD

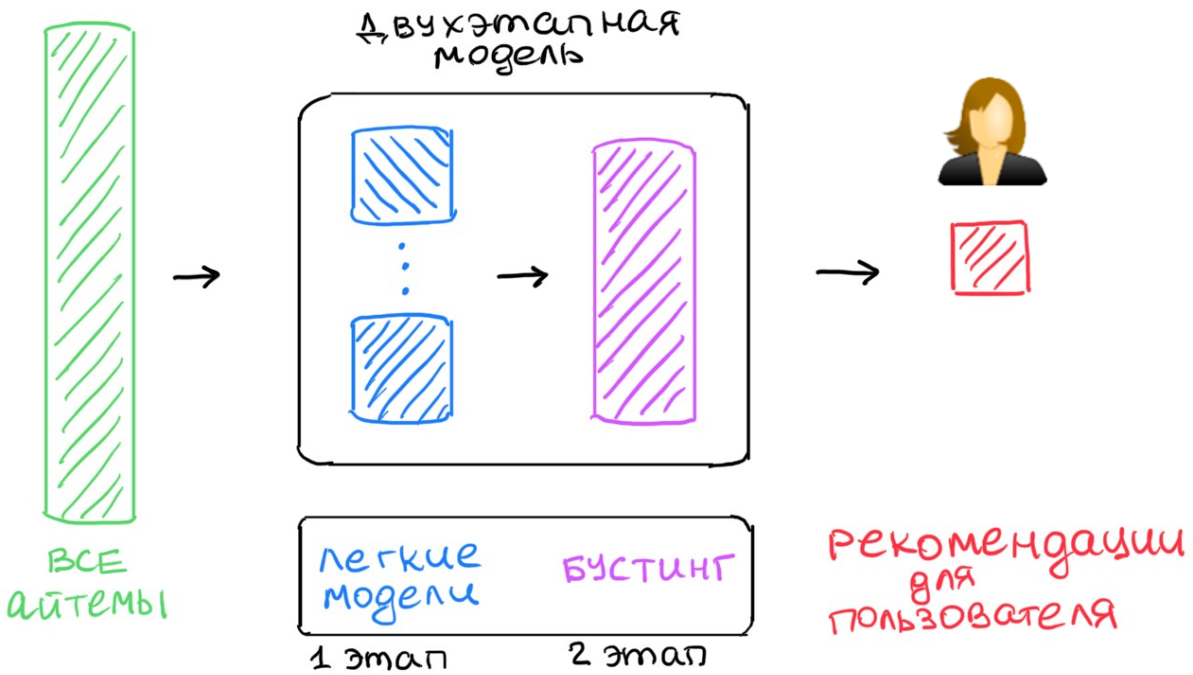
МТС ods.ai

Какие проблемы у простых моделей?

- учет временной составляющей
- линейность
- признаки, ранки, эмбеддинги
- как блендить результаты

Двухэтапная модель: основной принцип

Этап предикта



1 этап:
кандидаты

Отбор кандидатов: мотивация

Вычислительные **ресурсы ограничены** =

если в качестве **кандидатов** использовать **всё множество** доступных **айтемов**,

то через **модель 2 этапа** пройдут **миллионы** объектов

Кандидаты

- **Выход** более легкой и быстрой **модели**
  - LightFM, implicit KNN
- **Топ популярного**
  - по категориям / жанрам / соцдему
- Другие **эвристики**
- **Блендинг всех кандидатов**

2 этап:
ранжирование кандидатов
бустингом

Модель 2 этапа

Обучение на **парах (юзер, айтем)**
Цель: переранжировать кандидатов 1 этапа

- **Binary classification**
  [target = 0/1, Logloss]
- **Learning to rank**
  [target = (1,0), YetiRank CatBoost]
- **Regression**
  [target = R, RMSE]

в этой лекции будем рассматривать бинарную классификацию

Модель 2 этапа
Задача бинарной классификации

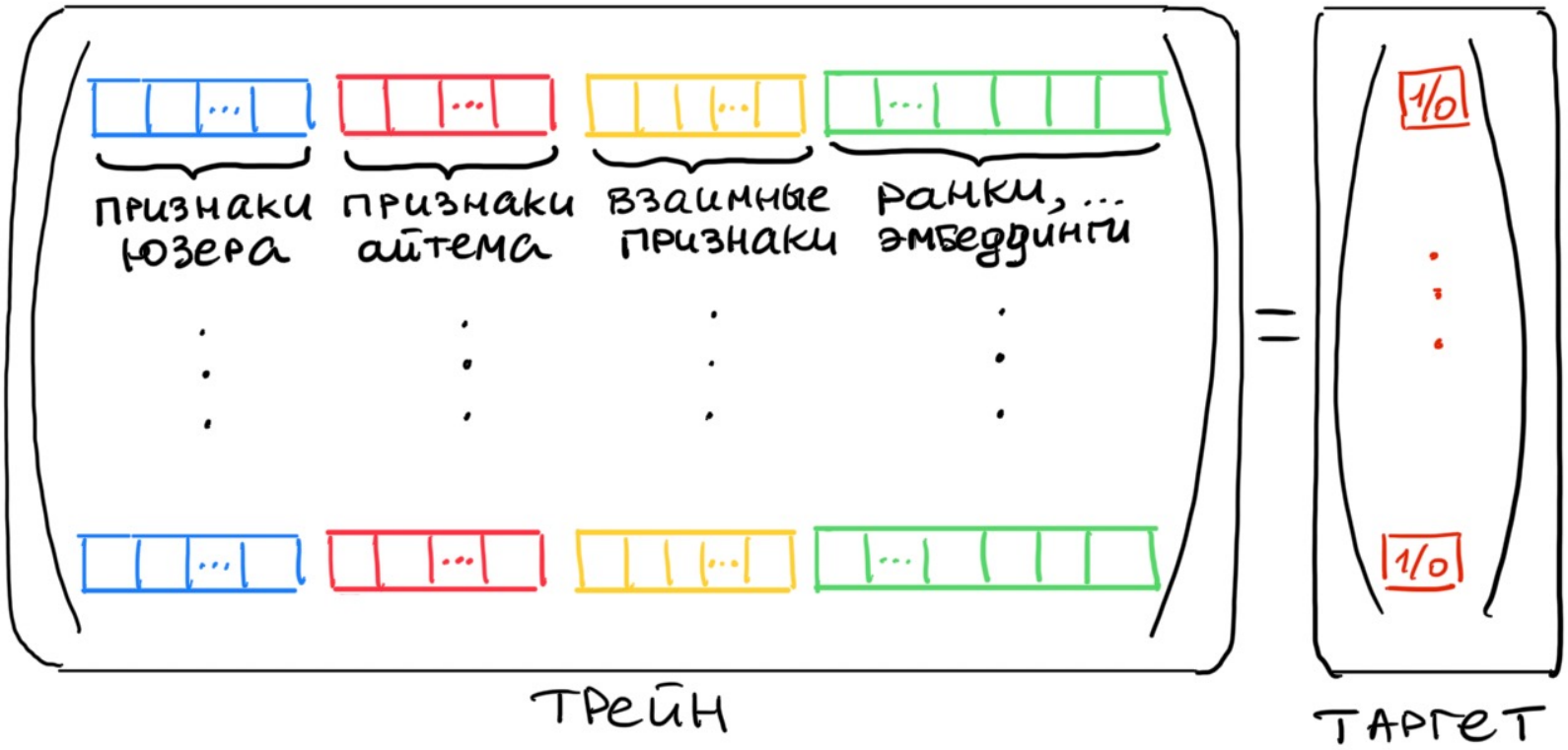
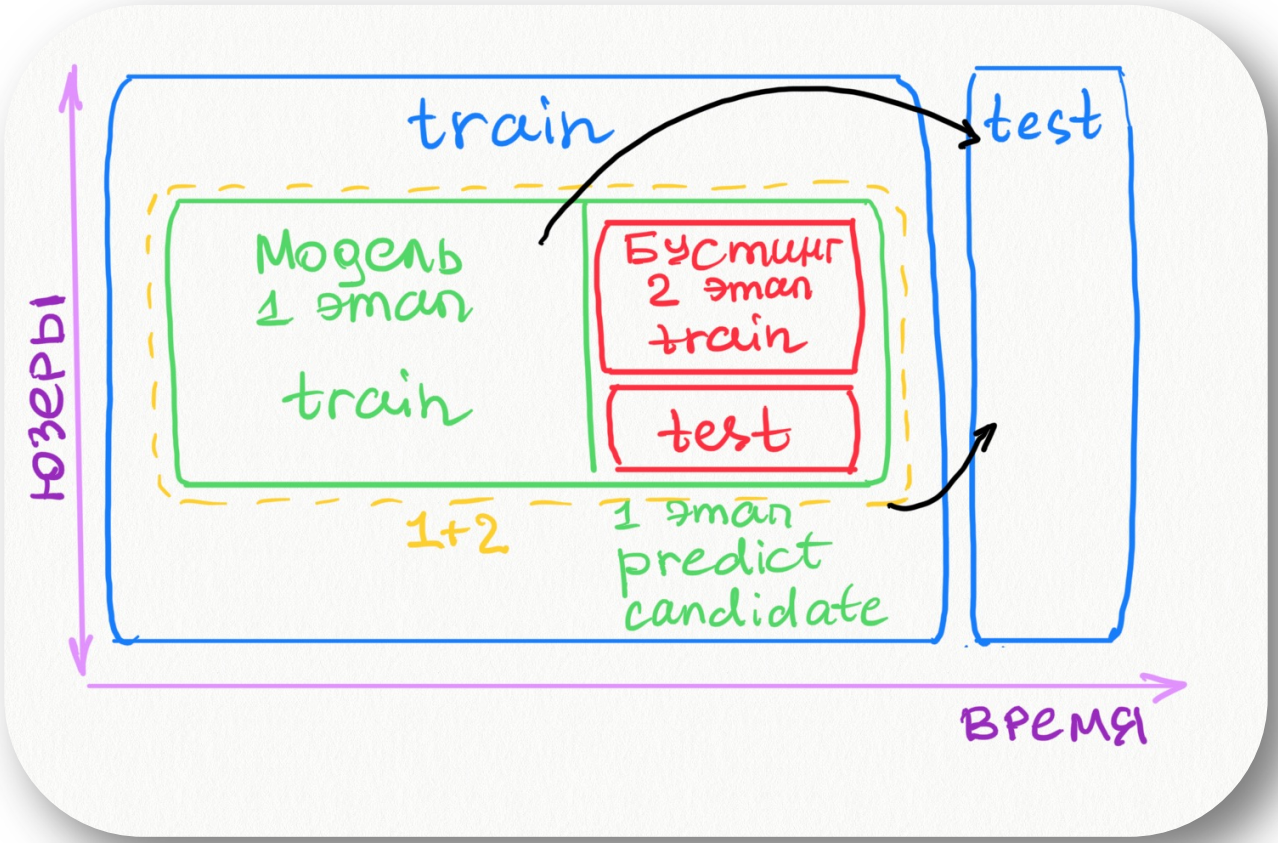


Схема валидации



Собираем трейн

Позитивные объекты ➕

(единички, “1”)

= взаимодействия пользователь – айтем

+ есть дата взаимодействия

Например,

1 = (юзер + трек, который он слушал)

Гиперпараметры:

- **Количество**
- **Период времени**
- **Фильтры**
  - только взаимодействия
    с айтемами-кандидатами из 1 этапа

Негативные объекты

(нолики, “0”)

= сэмпл из айтемов, с которым пользователь не взаимодействовал
+ нет даты взаимодействия

Например,

0 = (юзер + все треки, которые он не слушал)

Гиперпараметры:

- **Количество (размер сэмпла)**
- **Фильтры**
  - Только сэмпл из айтемов-кандидатов из 1 этапа
  - Сэмплирование: равномерное или по популярности

МТС ods.ai

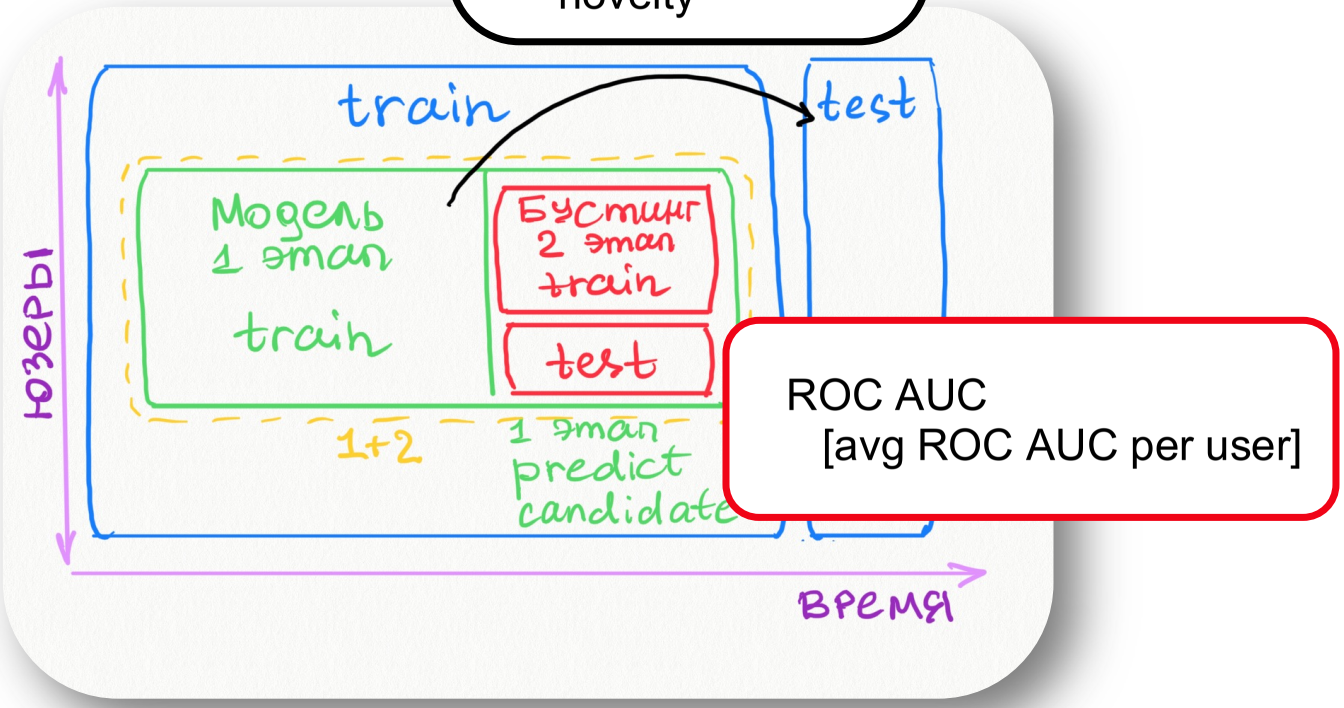
Признаки

- Фичи **айтемов**
- Фичи **юзеров**
- **Взаимные фичи** юзеров и айтемов
  “популярность” юзеров и “популярность” айтемов
- **Ранки, скоры, эмбеддинги** от легких моделей

Метрики качества

Классика
- precision@k
- recall@k
- MAP@k

Beyond accuracy
- diversity
- novelty



ROC AUC
[avg ROC AUC per user]

Особенности подхода

- Правильная валидация важна
- Правильный сбор трейна, эксперименты
- Много данных = проблема с ресурсами:
  - Если долго обучается модель 2 этапа
    - переобучаем раз в неделю, а применяем каждый день
    - переобучаем каждый день только модель для кандидатов 1 этапа
  - Долгий этап применения
    - Батчевый подход
    - Spark ML

Полезные ссылки

- Статья A Hybrid Approach to Music Playlist Continuation Based on Playlist-song Membership. Andreu Vall, Matthias Dorfer, Markus Schedl, and Gerhard Widmer. 2018
- Как мы решали задачу продолжения плейлистов на RecSys Challenge и заняли 3 место Статья на Хабре Статья на ACM
- Статья Rekko Challenge 2019 — как занять 2-е место в конкурсе по созданию рекомендательных систем
- Доклад Next-level recommendations: как сделать модель второго уровня в рекомендациях, Михаил Каменщиков
- Статья Neural Networks for YouTube Recommendations. Paul Covington, Jay Adams, Emre Sargin, 2016.

Спасибо за внимание!

Елисова Ирина

 @elisovaira

uplift статьи **хабр**
scikit–uplift **библиотека**

linkedin